

# Logit stick-breaking priors for Bayesian density regression

Tommaso Rigon

Department of Decision Sciences, Bocconi University  
and

Daniele Durante

Department of Statistical Sciences, University of Padova

March 27, 2017

## Abstract

There is an increasing focus in several fields on learning how the distribution of a response variable changes with a set of predictors. Bayesian nonparametric dependent mixture models provide a useful approach to flexibly address this goal, however many representations are characterized by difficult interpretation and intractable computational methods. Motivated by these issues, we describe a flexible class of predictor-dependent infinite Gaussian mixture models, which relies on a formal characterization of the stick-breaking construction via a continuation–ratio logistic regression, within an exponential family representation. We study the theoretical properties, and leverage this result to derive analytically three computational methods of routine use in Bayesian inference, covering simple Markov Chain Monte Carlo via Gibbs sampling, the Expectation Maximization algorithm, and a variational Bayes procedure for scalable inference. The algorithms associated with these methods are made available online at <https://github.com/tommasorigon/DLSBP>. We additionally compare the three computational strategies in an application to the Old Faithful Geyser dataset.

*Keywords:* Bayesian density regression; Continuation–ratio logistic regression; EM algorithm; Gibbs sampling; Variational Bayes.

# 1 Introduction

There is a growing interest in density regression methods which allow the entire distribution of a univariate response variable  $y \in \mathcal{Y}$  to be unknown, and changing with a vector of predictors  $\mathbf{x} \in \mathcal{X}$ . The increased flexibility provided by these procedures allows relevant improvements in inference and prediction compared to classical regression frameworks, as seen in applications to epidemiology (e.g. Dunson & Park 2008), meteorology (e.g. Gutiérrez et al. 2016), neuroscience (e.g. Wade et al. 2014), image analysis (e.g. Ren et al. 2011) and finance (e.g. Griffin & Steel 2011) — among others.

There is a wide set of alternative methodologies to provide flexible inference for conditional distributions, within a Bayesian nonparametric framework. Most of these methods represent generalizations of the marginal density estimation problem for  $f(y)$ , which is commonly addressed via Bayesian nonparametric mixture models of the form  $f(y) = \int_{\Theta} K(y; \boldsymbol{\theta}) dP(\boldsymbol{\theta})$ , where  $K(y; \boldsymbol{\theta})$  is a known parametric kernel indexed by  $\boldsymbol{\theta} \in \Theta$ , and  $P(\boldsymbol{\theta})$  denotes an unknown mixing measure which is assigned a flexible prior  $\Pi$ . Popular choices for  $\Pi$  are the Dirichlet process (Ferguson 1973, 1974, Sethuraman 1994), the two-parameter Poisson–Dirichlet process (Pitman & Yor 1997), and other almost surely discrete random measures having a stick-breaking representation (Ishwaran & James 2001). This choice leads to an infinite mixture model representation for  $f(y)$  of the form

$$f(y) = \int_{\Theta} K(y; \boldsymbol{\theta}) dP(\boldsymbol{\theta}) = \sum_{h=1}^{+\infty} \pi_h K(y; \boldsymbol{\theta}_h), \quad \pi_h = \nu_h \prod_{l=1}^{h-1} (1 - \nu_l), \quad h = 1, \dots, +\infty, \quad (1)$$

with  $\boldsymbol{\theta}_h \sim P_0$ , independently for  $h = 1, \dots, +\infty$ , and the stick-breaking weights  $\nu_h$ ,  $h = 1, \dots, +\infty$ , having independent  $\text{Beta}(a_h, b_h)$  priors, so that  $\sum_{h=1}^{+\infty} \pi_h = 1$  almost surely. Fixing  $a_h = 1$  and  $b_h = \alpha$  leads to a Dirichlet process mixture model, whereas the two-parameter Poisson–Dirichlet process mixture can be obtained by letting  $a_h = 1 - a$  and  $b_h = b + ha$ , with  $0 \leq a < 1$  and  $b > -a$ . Model (1) has computational benefits in allowing the implementation of simple Markov Chain Monte Carlo methods for inference (e.g. Escobar & West 1995, Neal 2000), and has been shown to provide a consistent procedure for density estimation (e.g. Ghosal et al. 1999, Tokdar 2006, Ghosal & Van Der Vaart 2007).

These results have motivated different generalizations of (1) to incorporate the conditional density inference problem for  $f(y | \mathbf{x})$ . In addressing this goal, a class of procedures

focus on modeling the joint density  $f(y, \mathbf{x})$  via Bayesian nonparametric mixtures of multivariate kernels, to induce a flexible posterior distribution for the conditional density of  $y$  given  $\mathbf{x}$  (Müller et al. 1996, Müller & Quintana 2010, Hannah et al. 2011). As discussed in Wade et al. (2014) these contributions may face computational and practical issues when the predictor space  $\mathcal{X}$  is large and complex due to the need to model the marginal density  $f(\mathbf{x})$ , which is effectively a nuisance quantity when the focus is on conditional inference. This result has motivated alternative methodologies explicitly focused on modeling  $f(y | \mathbf{x})$  via a generalization of (1) which allows the unknown random mixing measure  $P_{\mathbf{x}}(\boldsymbol{\theta})$  to change with  $\mathbf{x} \in \mathcal{X}$ , under a dependent stick-breaking characterization (MacEachern 1999, 2000). Popular representations consider predictor-independent mixing weights  $\pi_h$ ,  $h = 1, \dots, +\infty$  and incorporate changes with  $\mathbf{x} \in \mathcal{X}$  in the atoms  $\boldsymbol{\theta}_h(\mathbf{x})$ , for  $h = 1, \dots, +\infty$  (e.g. De Iorio et al. 2004, Gelfand et al. 2005, Caron et al. 2006, De la Cruz-Mesía et al. 2007).

Although the above models have been successfully applied in different contexts, covering ANOVA-type formulations, spatial statistics, time-series analysis and classification, as noted in MacEachern (2000) and Griffin & Steel (2006), the predictor-independent assumption for the mixing weights can have limited flexibility in modeling  $f(y | \mathbf{x})$ . This has motivated more general formulations allowing also  $\pi_h(\mathbf{x})$ ,  $h = 1, \dots, +\infty$ , to change with the predictors. Relevant examples include the order-based dependent Dirichlet process (Griffin & Steel 2006), the kernel stick-breaking process (Dunson & Park 2008), the infinite mixture models with predictor-dependent normalized weights (Antoniano-Villalobos et al. 2014), and recent representations for dynamic density estimation (Gutiérrez et al. 2016). These formulations represent a more broad class of priors for density regression and have appealing theoretical properties (Pati et al. 2013), however their flexibility comes at cost in terms of interpretation and computational tractability.

Motivated by the above discussion, we propose an alternative formulation to characterize changes in each mixing weight  $\pi_h(\mathbf{x})$ , with the covariates  $\mathbf{x} \in \mathcal{X}$ . This representation is carefully defined to provide simpler interpretation and improved computational tractability, while maintaining flexibility and theoretical support. We accomplish these goals via a simple logit stick-breaking construction which relates each  $\nu_h(\mathbf{x}) \in (0, 1)$  to a function of the covariates  $\eta_h(\mathbf{x}) \in \mathbb{R}$ , using the logistic link. Our contribution is closely related to

the probit stick-breaking prior of Rodriguez & Dunson (2011) leveraging the probit link function instead of the logistic one. However, as we will outline in the subsequent sections, our mapping can be formally interpreted as the canonical link for the continuation–ratio representation (Tutz 1991) of the hierarchical mechanism assigning units to mixture components, under the stick-breaking construction of the mixing weights. Our logistic mapping is also intimately related to the hierarchical mixtures of experts (Jordan & Jacobs 1994, Bishop & Svensén 2003), providing building-block results to implement scalable algorithms for estimation and approximate inference via the Expectation Maximization algorithm and variational Bayes, which are not provided in Rodriguez & Dunson (2011).

Ren et al. (2011) noticed a similar connection in their logistic stick-breaking process; however the focus is exclusively on nonparametric clustering of spatial and temporal data. Although we rely on a similar representation, our contribution is designed for more general density regression settings and provides additional results in terms of interpretation, computational implementation and theoretical support. For example, we show that our logit stick-breaking construction can be interpreted as continuation–ratio logistic regression for the assignment of the units to the mixture components. Leveraging the Pòlya-Gamma data augmentation for Bayesian logistic regression (Polson et al. 2013), this result facilitates the implementation of a simple Gibbs sampler which converges to the exact posterior and avoids the approximations required in Ren et al. (2011).

The remainder of the paper is organized as follows. In Section 2 we describe the logit stick-breaking prior, along with its properties and the formal interpretation via continuation–ratio logistic regression. Section 3 provides detailed derivation of three algorithms of routine use in Bayesian density regression, covering Gibbs sampling, the Expectation Maximization algorithm, and a variational Bayes approach for scalable inference. The performance of these methods is assessed in Section 4 with an application to the Old Faithful Geyser dataset. Concluding remarks are given in Section 5.

## 2 The logit stick-breaking prior

This section presents a formal construction of the logit stick-breaking prior via a continuation–ratio parameterization of the hierarchical mechanism assigning the units to mixture compo-

nents. Although our logit stick-breaking representation for the predictor–dependent mixing weights and the associated computational procedures apply to a wide set of dependent mixture models and kernels, we will mainly focus — for the sake of clarity — on a general class of predictor–dependent infinite Gaussian mixture models of the form

$$f(y \mid \mathbf{x}) = \int \frac{1}{\sigma} \phi \left\{ \frac{y - \boldsymbol{\lambda}(\mathbf{x})^\top \boldsymbol{\beta}}{\sigma} \right\} dP_{\mathbf{x}}(\boldsymbol{\beta}, \sigma) = \sum_{h=1}^{+\infty} \pi_h(\mathbf{x}) \frac{1}{\sigma_h} \phi \left\{ \frac{y - \boldsymbol{\lambda}(\mathbf{x})^\top \boldsymbol{\beta}_h}{\sigma_h} \right\}, \quad (2)$$

with  $\pi_h(\mathbf{x}) = \nu_h(\mathbf{x}) \prod_{l=1}^{h-1} \{1 - \nu_l(\mathbf{x})\}$ , and  $\boldsymbol{\beta}_h = (\beta_{1h}, \dots, \beta_{Ph})^\top$  a vector of coefficients linearly related to selected functions of the predictors  $\lambda_1(\mathbf{x}), \dots, \lambda_P(\mathbf{x})$ , comprising the vector  $\boldsymbol{\lambda}(\mathbf{x})$ . Formulation (2) is arguably the most widely used in Bayesian nonparametric density regression and has been shown to provide consistent estimates of  $f(y \mid \mathbf{x})$  in asymptotic settings (Pati et al. 2013), thereby motivating an in depth study of the associated properties and computational methods. Generalizations to other kernels will be also discussed.

## 2.1 Logit stick-breaking for infinite Gaussian mixture models

To provide a constructive representation of the logit stick-breaking prior, let us consider an equivalent formulation of the predictor–dependent mixture model in (2). In particular — following standard hierarchical representations of mixture models — independent samples  $y_1, \dots, y_n$  from the variable with density function factorized in (2), can be obtained from

$$(y_i \mid G_i = h, \boldsymbol{\theta}_h) \sim N\{\boldsymbol{\lambda}(\mathbf{x}_i)^\top \boldsymbol{\beta}_h, \sigma_h^2\}, \quad \text{pr}(G_i = h) = \pi_h(\mathbf{x}_i) = \nu_h(\mathbf{x}_i) \prod_{l=1}^{h-1} \{1 - \nu_l(\mathbf{x}_i)\}, \quad (3)$$

for every unit  $i = 1, \dots, n$ , with  $\boldsymbol{\theta}_h = (\boldsymbol{\beta}_h, \sigma_h^2) \sim P_0$  and  $G_i \in \{1, 2, \dots, +\infty\}$  an indicator variable denoting the mixture component associated with unit  $i$ . According to (3) each  $G_i$  has probability mass function  $p(G_i) = \prod_{h=1}^{+\infty} \pi_h(\mathbf{x}_i)^{\mathbb{1}(G_i=h)}$ , which can be rewritten — under the stick-breaking factorization for  $\pi_h(\mathbf{x}_i)$  in (3) — as

$$\begin{aligned} & \nu_1(\mathbf{x}_i)^{\mathbb{1}(G_i=1)} \{1 - \nu_1(\mathbf{x}_i)\}^{1-\mathbb{1}(G_i=1)} \dots \nu_h(\mathbf{x}_i)^{\mathbb{1}(G_i=h)} \{1 - \nu_h(\mathbf{x}_i)\}^{\mathbb{1}(G_i>h-1)-\mathbb{1}(G_i=h)} \dots \\ &= \prod_{h=1}^{+\infty} \exp[\mathbb{1}(G_i = h) \log[\nu_h(\mathbf{x}_i)/\{1 - \nu_h(\mathbf{x}_i)\}] + \mathbb{1}(G_i > h - 1) \log\{1 - \nu_h(\mathbf{x}_i)\}]. \end{aligned} \quad (4)$$

Hence, the distribution of each component membership indicator  $G_i$  can be factorized as the product of conditionally independent Bernoulli probability mass functions for the binary

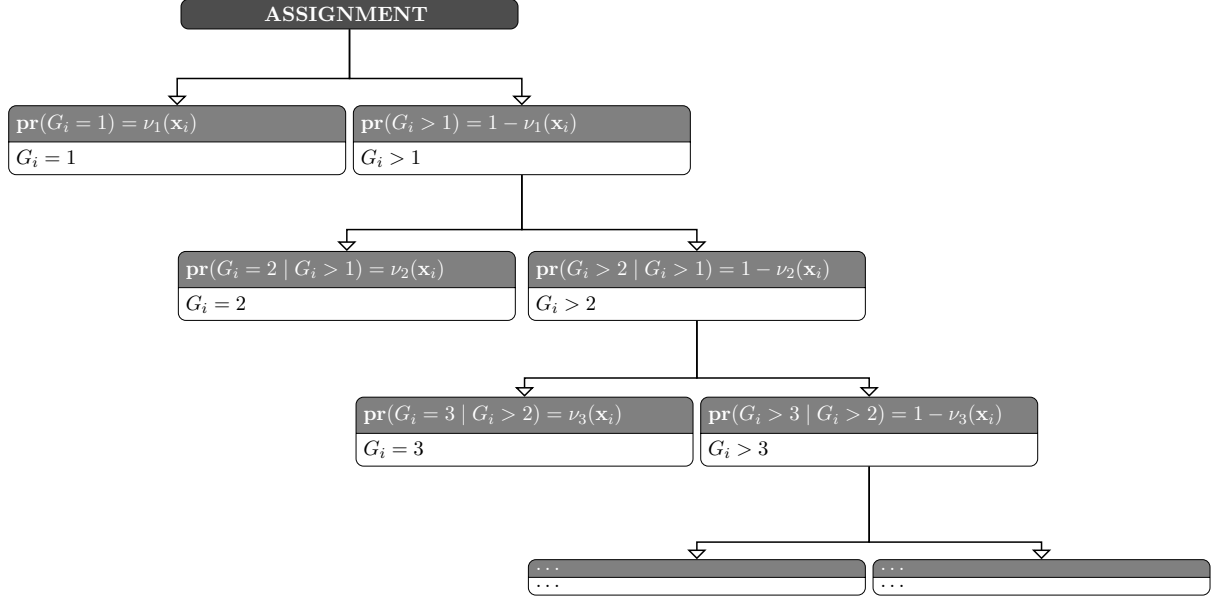


Figure 1: Representation of the sequential mechanism to sample  $G_i$ .

variables  $\{\mathbb{1}(G_i = h) \mid \mathbb{1}(G_i > h - 1) = 1\} \sim \text{Bern}\{\nu_h(\mathbf{x}_i) = \text{pr}(G_i = h \mid G_i > h - 1)\}$ , for  $h = 1, \dots, +\infty$ , having natural parameters  $\eta_h(\mathbf{x}_i) = \log[\nu_h(\mathbf{x}_i)/\{1 - \nu_h(\mathbf{x}_i)\}] \in \mathfrak{R}$  and logistic canonical link. This result provides support for our logit stick-breaking factorization

$$\pi_h(\mathbf{x}_i) = \nu_h(\mathbf{x}_i) \prod_{l=1}^{h-1} \{1 - \nu_l(\mathbf{x}_i)\} = \frac{1}{1 + \exp\{-\eta_h(\mathbf{x}_i)\}} \prod_{l=1}^{h-1} \left[ 1 - \frac{1}{1 + \exp\{-\eta_l(\mathbf{x}_i)\}} \right], \quad (5)$$

for every  $h = 1, \dots, +\infty$ , while allowing simple interpretation of the stick-breaking construction via a continuation–ratio logistic regression (Tutz 1991), described in Figure 1.

In particular, in the first step of this continuation–ratio generative mechanism, unit  $i$  is either assigned to the first component with probability  $\nu_1(\mathbf{x}_i) = [1 + \exp\{-\eta_1(\mathbf{x}_i)\}]^{-1}$  or to one of the others with complement probability. If  $G_i = 1$  the process stops, otherwise it continues considering the reduced space  $\{2, \dots, +\infty\}$ . A generic step  $h$  is reached if  $i$  has not been assigned to  $1, \dots, h - 1$ , and the decision at the  $h$ th step will be to either allocate  $i$  to component  $h$  with probability  $\nu_h(\mathbf{x}_i) = [1 + \exp\{-\eta_h(\mathbf{x}_i)\}]^{-1}$  or to one of the subsequent components  $h + 1, \dots, +\infty$  with probability  $1 - \nu_h(\mathbf{x}_i)$ , conditioned on  $G_i \in \{h, \dots, +\infty\}$ . This generative mechanism plays a key role in developing simple computational procedures.

To conclude our Bayesian representation we require priors for the parameters  $\eta_h(\mathbf{x}_i) \in \mathfrak{R}$ , characterizing the log-odds of each conditional probability  $\nu_h(\mathbf{x}_i) \in (0, 1)$ ,  $h = 1, \dots, +\infty$

in the continuation–ratio logistic regressions. A natural choice — consistent with classical generalized linear models representations (e.g. McCullagh & Nelder 1989) — is to define the log-odds as a linear combination of selected functions  $\boldsymbol{\psi}(\mathbf{x}_i) = \{\psi_1(\mathbf{x}_i), \dots, \psi_R(\mathbf{x}_i)\}^\top$  of the covariates and consider Gaussian priors for the coefficients, obtaining

$$\text{logit}\{\nu_h(\mathbf{x}_i)\} = \eta_h(\mathbf{x}_i) = \boldsymbol{\psi}(\mathbf{x}_i)^\top \boldsymbol{\alpha}_h = \sum_{r=1}^R \alpha_{rh} \psi_r(\mathbf{x}_i), \quad \boldsymbol{\alpha}_h \sim N_R(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha), \quad h = 1, \dots, +\infty. \quad (6)$$

Although the linearity assumption in equation (6) may seem restrictive, it is worth noticing that flexible formulations for  $\eta_h(\mathbf{x}_i)$ , including regression via splines and Gaussian processes, induce relations that are linear in the coefficients. Moreover, as we will outline in Section 3, the linearity assumption greatly simplifies computations, while inducing a logistic-normal prior for each  $\nu_h(\mathbf{x}_i)$ ,  $h = 1, \dots, +\infty$ , with well defined moments (Aitchison & Shen 1980). Hence, the logit stick-breaking does not induce Beta distributed stick-breaking weights, and therefore cannot be included in the general class of stick-breaking priors discussed in Ishwaran & James (2001). However, as outlined in Section 2.2, many relevant properties characterizing the priors discussed in Ishwaran & James (2001) are met also under our case.

## 2.2 Properties of the logit stick-breaking prior

Let  $\Theta$  be a complete and separable metric space endowed with the Borel  $\sigma$ -algebra  $\mathcal{B}(\Theta)$ , and let  $\{P_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$  denote the class of predictor–dependent random probability measures on  $\Theta$ , induced by our logit stick-breaking prior via

$$P_{\mathbf{x}}(\cdot) = \sum_{h=1}^{+\infty} \pi_h(\mathbf{x}) \delta_{\boldsymbol{\theta}_h}(\cdot), \quad \pi_h(\mathbf{x}) = \nu_h(\mathbf{x}) \prod_{l=1}^{h-1} \{1 - \nu_l(\mathbf{x})\}, \quad \text{logit}\{\nu_h(\mathbf{x})\} = \boldsymbol{\psi}(\mathbf{x}_i)^\top \boldsymbol{\alpha}_h, \quad (7)$$

with independent and identically distributed atoms  $\boldsymbol{\theta}_h \sim P_0$ ,  $h = 1, \dots, +\infty$  from the space  $\{\Theta, \mathcal{B}(\Theta)\}$ , and  $\boldsymbol{\alpha}_h \sim N_R(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha)$  for every  $h = 1, \dots, +\infty$ . As discussed in Section 2.1, representation (7) does not provide Beta distributed priors for the stick-breaking weights  $\nu_h(\mathbf{x})$ ,  $h = 1, \dots, +\infty$ . However, in line with the random measures outlined in Ishwaran & James (2001), also our logit stick-breaking prior provides a well defined predictor–dependent random probability measure  $P_{\mathbf{x}}$  at every  $\mathbf{x} \in \mathcal{X}$ , as discussed in Proposition 1.

**Proposition 1.** *For any  $\mathbf{x} \in \mathcal{X}$ ,  $\sum_{h=1}^{+\infty} \pi_h(\mathbf{x}) = 1$  almost surely, with  $\pi_h(\mathbf{x})$  factorized as in (7) and  $\boldsymbol{\alpha}_h \sim N_R(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha)$  for every  $h = 1, \dots, +\infty$ .*

**Proof:** Following Lemma 1 in Ishwaran & James (2001), we have that  $\sum_{h=1}^{+\infty} \pi_h(\mathbf{x}) = 1$  almost surely if and only if  $\sum_{h=1}^{+\infty} \mathbb{E}[\log\{1 - \nu_h(\mathbf{x})\}] = -\infty$ . Since  $\log\{1 - \nu_h(\mathbf{x})\}$  is concave in  $\nu_h(\mathbf{x})$  for every  $\mathbf{x} \in \mathcal{X}$  and  $h = 1, \dots, +\infty$ , by Jensen inequality we have  $\mathbb{E}[\log\{1 - \nu_h(\mathbf{x})\}] \leq \log[1 - \mathbb{E}\{\nu_h(\mathbf{x})\}]$ . Hence, since  $\nu_h(\mathbf{x}) \in (0, 1)$ , from the usual properties of the expectation we have that  $0 < \mathbb{E}\{\nu_h(\mathbf{x})\} = \mu_{1\nu}(\mathbf{x}) < 1$ , thereby providing  $\log\{1 - \mu_{1\nu}(\mathbf{x})\} < 0$ . Therefore,  $\sum_{h=1}^{+\infty} \mathbb{E}[\log\{1 - \nu_h(\mathbf{x})\}] \leq \sum_{h=1}^{+\infty} \log\{1 - \mu_{1\nu}(\mathbf{x})\} = -\infty$ , proving Proposition 1.  $\square$

Although we focus on the infinite case, our logit stick-breaking prior is well defined also in truncated models considering a finite number of mixture components  $H$ . Consistent with Ishwaran & James (2001), in this case it suffices to model the first  $H - 1$  weights  $\nu_1(\mathbf{x}), \dots, \nu_{H-1}(\mathbf{x})$  and let  $\nu_H(\mathbf{x}) = 1$  for any  $\mathbf{x} \in \mathcal{X}$ , to ensure condition  $\sum_{h=1}^H \pi_h(\mathbf{x}) = 1$ .

Results in Proposition 1 motivate further analyses of the logit stick-breaking prior. In particular, consistent with theoretical studies on other stick-breaking priors not belonging to the class discussed in Ishwaran & James (2001) — e.g. Dunson & Park (2008), Rodriguez & Dunson (2011) — Proposition 2 provides additional insights on the moments of the predictor-dependent random probability measure induced by our logit stick-breaking prior.

**Proposition 2.** *For every  $\mathbf{x} \in \mathcal{X}$  and  $B \in \mathcal{B}(\Theta)$  the expectation of  $P_{\mathbf{x}}(B)$  is  $\mathbb{E}\{P_{\mathbf{x}}(B)\} = P_0(B)$ , whereas the variance of  $P_{\mathbf{x}}(B)$  for any truncated version of  $P_{\mathbf{x}}(\cdot)$  in (7) with  $H > 1$  components — including the infinite case — is*

$$\text{var}\{P_{\mathbf{x}}(B)\} = P_0(B)\{1 - P_0(B)\} \frac{\mu_{2\nu}(\mathbf{x})\{1 - [1 - 2\mu_{1\nu}(\mathbf{x}) + \mu_{2\nu}(\mathbf{x})]^H\}}{2\mu_{1\nu}(\mathbf{x}) - \mu_{2\nu}(\mathbf{x})},$$

where  $\mu_{1\nu}(\mathbf{x}) = \mathbb{E}\{\nu_h(\mathbf{x})\}$  and  $\mu_{2\nu}(\mathbf{x}) = \mathbb{E}\{\nu_h(\mathbf{x})^2\}$  for every  $h = 1, \dots, +\infty$ . The covariance at two different predictor values  $\mathbf{x} \in \mathcal{X}$  and  $\mathbf{x}' \in \mathcal{X}$ ,  $\mathbf{x} \neq \mathbf{x}'$ , is instead

$$\text{cov}\{P_{\mathbf{x}}(B), P_{\mathbf{x}'}(B)\} = P_0(B)\{1 - P_0(B)\} \frac{\mu_{2\nu}(\mathbf{x}, \mathbf{x}')\{1 - [1 - \mu_{1\nu}(\mathbf{x}) - \mu_{1\nu}(\mathbf{x}') + \mu_{2\nu}(\mathbf{x}, \mathbf{x}')]^H\}}{\mu_{1\nu}(\mathbf{x}) + \mu_{1\nu}(\mathbf{x}') - \mu_{2\nu}(\mathbf{x}, \mathbf{x}')},$$

with  $\mu_{2\nu}(\mathbf{x}, \mathbf{x}') = \mathbb{E}\{\nu_h(\mathbf{x})\nu_h(\mathbf{x}')\}$ .

**Proof:** Results are a direct consequence of the calculations in Appendix 2 and Appendix 6 in Rodriguez & Dunson (2011), after replacing the probit link with the logistic one.  $\square$

According to Proposition 2, the expectation of  $P_{\mathbf{x}}(\cdot)$  coincides with the base measure  $P_0(\cdot)$  which can be therefore interpreted as the prior guess for the mixing measure at any  $\mathbf{x} \in \mathcal{X}$ . This quantity is predictor-independent, meaning that a priori we are not forcing



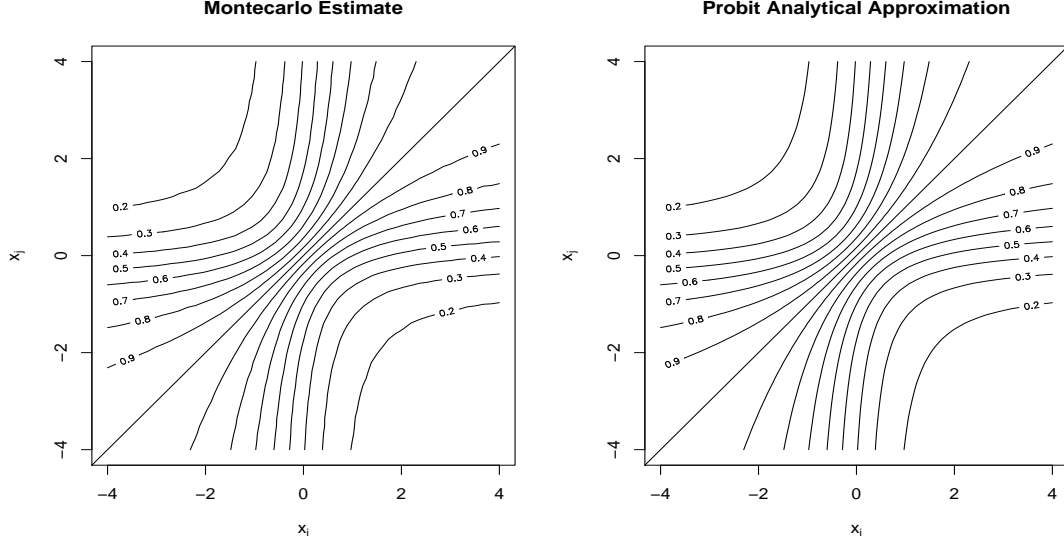


Figure 2: Left: Contour plot of a Monte Carlo estimate for the correlation between  $P_{\mathbf{x}}(B)$  and  $P_{\mathbf{x}'}(B)$ , at the infinite case, with  $0 < P_0(B) < 1$ ,  $\eta_h(\mathbf{x}) = \alpha_{1h} + \alpha_{2h}x$  and  $\boldsymbol{\alpha}_h = (\alpha_{1h}, \alpha_{2h})^\top \sim N_2(0, 10^4 I_2)$ , for values  $x, x'$  in  $(-4, 4)$ . Right: Same quantity, relying on the analytical approximation of the logistic link.

particular dependence structure between the atoms  $\boldsymbol{\theta}$  and the predictors. The variance changes instead with the predictors via a function of the first two moments of the logistic-normal stick-breaking weights. Note that, since each  $\nu_h(\mathbf{x})$  is bounded between 0 and 1, we have  $\nu_h(\mathbf{x}) \geq \nu_h(\mathbf{x})^2$  for every  $h = 1, \dots, +\infty$  and  $\mathbf{x} \in \mathcal{X}$ , implying  $0 < \mu_{2\nu}(\mathbf{x}) \leq \mu_{1\nu}(\mathbf{x}) < 1$ . These results provide the bound  $1 - 2\mu_{1\nu}(\mathbf{x}) + \mu_{2\nu}(\mathbf{x}) < 1$ , which leads to a well defined limiting variance for the infinite case  $H \rightarrow +\infty$  equal to  $P_0(B)\{1 - P_0(B)\}\mu_{2\nu}(\mathbf{x})\{2\mu_{1\nu}(\mathbf{x}) - \mu_{2\nu}(\mathbf{x})\}^{-1}$ . The limiting covariance is instead  $P_0(B)\{1 - P_0(B)\}\mu_{2\nu}(\mathbf{x}, \mathbf{x}')\{\mu_{1\nu}(\mathbf{x}) + \mu_{1\nu}(\mathbf{x}') - \mu_{2\nu}(\mathbf{x}, \mathbf{x}')\}^{-1}$ , after noticing that  $\mu_{1\nu}(\mathbf{x}) \geq \mu_{2\nu}(\mathbf{x}, \mathbf{x}')$ ,  $\mu_{1\nu}(\mathbf{x}') \geq \mu_{2\nu}(\mathbf{x}, \mathbf{x}')$  and  $1 - \mu_{1\nu}(\mathbf{x}) - \mu_{1\nu}(\mathbf{x}') + \mu_{2\nu}(\mathbf{x}, \mathbf{x}') < 1$ . Hence the association is always positive and increases the closer  $\mathbf{x}$  is to  $\mathbf{x}'$ . This behavior is illustrated in Figure 2.

Although results in Proposition 2 provide simple expressions for  $E\{P_{\mathbf{x}}(B)\}$ ,  $\text{var}\{P_{\mathbf{x}}(B)\}$  and  $\text{cov}\{P_{\mathbf{x}}(B), P_{\mathbf{x}'}(B)\}$ , their computation requires the moments of logistic-normal priors for the stick-breaking weights, induced by representation (6). Unfortunately these quantities are not available in explicit form (e.g. Aitchison & Shen 1980), however Proposition 3

provides a simple procedure to accurately approximate the moments of logit stick-breaking weights leveraging a connection with the probit stick-breaking priors.

**Proposition 3.** *The logit stick-breaking prior described in representation (6), can be accurately approximated by a probit stick-breaking process  $\nu_h(\mathbf{x}) \approx \Phi\{\boldsymbol{\psi}(\mathbf{x})^\top \bar{\boldsymbol{\alpha}}_h\}$ , with  $\bar{\boldsymbol{\alpha}}_h = \boldsymbol{\alpha}_h \sqrt{\pi/8} \sim N_R\{\sqrt{\pi/8}\boldsymbol{\mu}_\alpha, (\pi/8)\boldsymbol{\Sigma}_\alpha\}$ , for every  $\mathbf{x} \in \mathcal{X}$  and  $h = 1, \dots, +\infty$ .*

**Proof:** Consistent with results in Amemiya (1981), the logistic link  $\{1 + \exp(-\boldsymbol{\psi}(\mathbf{x})^\top \boldsymbol{\alpha}_h)\}^{-1}$  can be accurately approximated by  $\Phi\{\boldsymbol{\psi}(\mathbf{x})^\top \boldsymbol{\alpha}_h \sqrt{\pi/8}\}$ . Therefore

$$\nu_h(\mathbf{x}) = \{1 + \exp(-\boldsymbol{\psi}(\mathbf{x})^\top \boldsymbol{\alpha}_h)\}^{-1} \approx \Phi\{\boldsymbol{\psi}(\mathbf{x})^\top \boldsymbol{\alpha}_h \sqrt{\pi/8}\} = \Phi\{\boldsymbol{\psi}(\mathbf{x})^\top \bar{\boldsymbol{\alpha}}_h\}, \quad \mathbf{x} \in \mathcal{X},$$

with  $\bar{\boldsymbol{\alpha}}_h \sim \sqrt{\pi/8}N_R(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha)$ , for every  $h = 1, \dots, +\infty$ , concluding the proof.  $\square$

According to Proposition 3, the logit stick-breaking can be approximated by a probit stick-breaking process, up to a scale transformation of the prior for the coefficients  $\boldsymbol{\alpha}_h$ ,  $h = 1, \dots, +\infty$ . This result allows simple approximation for the moments of the logistic-normal priors on the stick-breaking weights by rescaling those provided in Rodriguez & Dunson (2011) for the probit stick-breaking. As shown in Figure 2, this analytical approximation provides indistinguishable results when compared to a Monte Carlo estimate, motivating the use of our computational algorithms also under the probit link. In fact, a researcher considering a probit stick-breaking process, could easily perform inference leveraging our algorithms, after rescaling the prior for the coefficients in the linear predictor by  $\sqrt{8/\pi}$ .

We conclude the analysis of the logit stick-breaking properties by studying how a truncated version of (7) approximates the infinite process. Although there are some computational methods for the infinite representation, these algorithms are not necessarily more tractable than those relying on a finite truncation, and still require approximations. In line with Rodriguez & Dunson (2011) and Ren et al. (2011), we develop detailed computational methods based on a finite representation, and discuss generalizations to the infinite case. This choice allows a more direct comparison between the algorithms proposed, and — according to Theorem 1 — provides an accurate approximation of the infinite representation.

**Theorem 1.** *For a sample  $\mathbf{y} = (y_1, \dots, y_n)^\top$  with covariates  $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}^\top$ , let*

$$f_H(\mathbf{y} \mid \mathbf{X}) = E_{P_{\mathbf{x}_i}^H} \left\{ \prod_{i=1}^n f_H(y_i \mid \mathbf{x}_i) \right\} = E_{P_{\mathbf{x}_i}^H} \left( \prod_{i=1}^n \left[ \int \frac{1}{\sigma} \phi \left\{ \frac{y_i - \boldsymbol{\lambda}(\mathbf{x}_i)^\top \boldsymbol{\beta}}{\sigma} \right\} dP_{\mathbf{x}_i}^H(\boldsymbol{\beta}, \sigma) \right] \right),$$

the expected joint density of the data based on a truncated version of the logit stick-breaking prior in (7) with  $H$  component, and let  $f_\infty(\mathbf{y} \mid \mathbf{X})$  be the same quantity in the infinite case. Then the total variation distance between  $f_H(\mathbf{y} \mid \mathbf{X})$  and  $f_\infty(\mathbf{y} \mid \mathbf{X})$  is bounded as follow

$$\|f_H(\mathbf{y} \mid \mathbf{X}) - f_\infty(\mathbf{y} \mid \mathbf{X})\|_1 \leq \sum_{i=1}^n \{1 - \mu_{1\nu}(\mathbf{x}_i)\}^{H-1}.$$

**Proof:** Following Ishwaran & James (2002) we can bound  $\|f_H(\mathbf{y} \mid \mathbf{X}) - f_\infty(\mathbf{y} \mid \mathbf{X})\|_1$  by

$$\|f_H(\mathbf{y} \mid \mathbf{X}) - f_\infty(\mathbf{y} \mid \mathbf{X})\|_1 \leq \mathbb{E} \left[ \prod_{i=1}^n \left\{ \left| \sum_{h=1}^H \pi_h(\mathbf{x}_i) \delta_{\beta_h, \sigma_h} - \sum_{h=1}^{+\infty} \pi_h(\mathbf{x}_i) \delta_{\beta_h, \sigma_h} \right| \right\} \right]$$

with  $\pi_H(\mathbf{x}_i) = 1 - \sum_{h=1}^{H-1} \pi_h(\mathbf{x}_i)$  to ensure that the stick-breaking prior is well defined in the truncated case. Focusing on  $\left| \sum_{h=1}^H \pi_h(\mathbf{x}_i) \delta_{\beta_h, \sigma_h} - \sum_{h=1}^{+\infty} \pi_h(\mathbf{x}_i) \delta_{\beta_h, \sigma_h} \right|$  we can easily obtain

$$\left| \sum_{h=1}^H \pi_h(\mathbf{x}_i) \delta_{\beta_h, \sigma_h} - \sum_{h=1}^{+\infty} \pi_h(\mathbf{x}_i) \delta_{\beta_h, \sigma_h} \right| = \left| \sum_{h=H}^{+\infty} \pi_h(\mathbf{x}_i) (\delta_{\beta_H, \sigma_H} - \delta_{\beta_h, \sigma_h}) \right| \leq \sum_{h=H}^{+\infty} \pi_h(\mathbf{x}_i) |\delta_{\beta_H, \sigma_H} - \delta_{\beta_h, \sigma_h}|,$$

which can be further bounded by  $\sum_{h=H}^{+\infty} \pi_h(\mathbf{x}_i)$ . Therefore, to prove Theorem 1 we need to compute  $\mathbb{E}[\prod_{i=1}^n \{\sum_{h=H}^{+\infty} \pi_h(\mathbf{x}_i)\}] \leq \mathbb{E}\{\sum_{i=1}^n \sum_{h=H}^{+\infty} \pi_h(\mathbf{x}_i)\} = \sum_{i=1}^n [1 - \sum_{h=1}^{H-1} \mathbb{E}\{\pi_h(\mathbf{x}_i)\}]$ , where  $\sum_{h=1}^{H-1} \mathbb{E}\{\pi_h(\mathbf{x}_i)\} = \sum_{h=1}^{H-1} \mu_{1\nu}(\mathbf{x}_i) \{1 - \mu_{1\nu}(\mathbf{x}_i)\}^{h-1} = 1 - \{1 - \mu_{1\nu}(\mathbf{x}_i)\}^{H-1}$ .  $\square$

According to Theorem 1, for fixed  $n$  and  $\mathbf{X}$ , the total variation distance between  $f_H(\mathbf{y} \mid \mathbf{X})$  and  $f_\infty(\mathbf{y} \mid \mathbf{X})$  vanishes as  $H \rightarrow +\infty$ , meaning that  $f_H(\mathbf{y} \mid \mathbf{X})$  converges in distribution to  $f_\infty(\mathbf{y} \mid \mathbf{X})$  when  $H \rightarrow +\infty$ . This rate of decay is exponential in  $H$ , and therefore — as for the Dirichlet process and the probit stick-breaking prior — the number of components has not to be very large in practice to accurately approximate the infinite representation.

### 3 Computational methods for Bayesian inference

This section provides detailed derivation of three computational methods for inference on the general class of predictor-dependent infinite Gaussian mixture models outlined in (2), with logit stick-breaking prior (6) for the mixing weights. In particular we consider a Gibbs sampler which converges to the exact posterior, an Expectation Maximization algorithm for fast estimation, and a variational Bayes approximation for scalable inference.

These methods exploit the hierarchical representation of model (2), which is described in equation (3), along with the continuation-ratio characterization of the logit stick-breaking

prior discussed in Section 2. In fact, conditioned on the component membership indicators  $\mathbf{G} = (G_1, \dots, G_n)$ , the model reduces to a set of separate Gaussian linear regressions — one for each mixture component — allowing inference for the kernel parameters  $\beta_h$  and  $\sigma_h$ , via standard methods when  $\beta_h \sim N_P(\mu_\beta, \Sigma_\beta)$  and  $\sigma_h^{-2} \sim \text{Ga}(a_\sigma, b_\sigma)$ ,  $h = 1, \dots, +\infty$ . Moreover, exploiting  $\mathbf{G} = (G_1, \dots, G_n)$ , and the continuation-ratio representation, inference for the logit stick-breaking parameters  $\alpha_h$ ,  $h = 1, \dots, +\infty$  in (6), proceeds as in a Bayesian logistic regression for the indicators  $\{\mathbb{1}(G_i = h) \mid G_i > h - 1\} \sim \text{Bern}[\nu_h(\mathbf{x}_i) = \{1 + \exp(-\psi(\mathbf{x})^\top \alpha_h)\}^{-1}]$ ,  $h = 1, \dots, +\infty$ . Adapting results from the recently developed Pòlya-Gamma data augmentation scheme (Polson et al. 2013) to our statistical model, this inference focus can be easily accomplished exploiting the following result:

$$\omega_{ih}^{-1}[\{\mathbb{1}(G_i = h) \mid G_i > h - 1\} - 0.5] \mid \omega_{ih}, \alpha_h \sim N\{\psi(\mathbf{x}_i)^\top \alpha_h, \omega_{ih}^{-1}\}, \quad (8)$$

$$\omega_{ih} \mid \alpha_h \sim \text{PG}\{1, \psi(\mathbf{x}_i)^\top \alpha_h\}, \quad (9)$$

independently for every  $i = 1, \dots, n$  and  $h = 1, \dots, +\infty$ , where  $\text{PG}\{1, \psi(\mathbf{x}_i)^\top \alpha_h\}$  denotes the Pòlya-Gamma random variable. Hence inference on the stick-breaking parameters, can be recast in terms of Bayesian linear regression with Gaussian transformed data  $[\{\mathbb{1}(G_i = h) \mid G_i > h - 1\} - 0.5]/\omega_{ih}$ , for  $i = 1, \dots, n$  and  $h = 1, \dots, +\infty$ . This result, along with the continuation-ratio representation, provide relevant benefits in the derivation of the different computational strategies, as we will outline in the following sections. Refer also to Choi & Hobert (2013) for theoretical properties on the Pòlya-Gamma data augmentation.

### 3.1 Gibbs sampler

In deriving the Gibbs sampler algorithm for the statistical model in (2), with logit stick-breaking prior defined in equation (6), we first focus on a dependent mixture of Gaussians with fixed  $H$ , and then discuss generalizations to the infinite representation.

Let  $\Lambda_h(\mathbf{x})$  and  $\Psi_h(\mathbf{x})$  denote the  $n_h \times P$  and the  $\bar{n}_h \times R$  predictor matrices in (2) and (6) having row entries  $\lambda(\mathbf{x}_i)^\top$  and  $\psi(\mathbf{x}_i)^\top$ , for only those statistical units  $i$  such that  $G_i = h$  and  $G_i > h - 1$ , respectively, the Gibbs sampler for the truncated representation of model (2) alternates between the full conjugate updating steps described in Algorithm 1. Note that step [1] can be run in parallel across units  $i = 1, \dots, n$ , whereas parallel computing for the different mixture components  $h = 1, \dots, H$  can be easily implemented in steps [2]–[4].

---

**Algorithm 1:** Steps of the Gibbs sampler for dependent finite mixture of Gaussians

---

[1] **for**  $i$  from 1 to  $n$  **do** update  $G_i$  from the discrete variable with probabilities

$$\text{pr}(G_i = h \mid -) = \frac{\left[ \nu_h(\mathbf{x}_i) \prod_{l=1}^{h-1} \{1 - \nu_l(\mathbf{x}_i)\} \right] \frac{1}{\sigma_h} \phi \left\{ \frac{y_i - \boldsymbol{\lambda}(\mathbf{x}_i)^\top \boldsymbol{\beta}_h}{\sigma_h} \right\}}{\sum_{q=1}^H \left[ \nu_q(\mathbf{x}_i) \prod_{l=1}^{q-1} \{1 - \nu_l(\mathbf{x}_i)\} \right] \frac{1}{\sigma_q} \phi \left\{ \frac{y_i - \boldsymbol{\lambda}(\mathbf{x}_i)^\top \boldsymbol{\beta}_q}{\sigma_q} \right\}},$$

for every  $h = 1, \dots, H$ .

[2] Update the parameters  $\boldsymbol{\alpha}_h$ ,  $h = 1, \dots, H - 1$ , for the logit stick-breaking prior in equation (6). Exploiting the continuation-ratio representation and the results from the Pòlya-Gamma data augmentation (8)–(9), this step proceeds as follows.

**for**  $h$  from 1 to  $H - 1$  **do** update the logit stick-breaking parameters  $\boldsymbol{\alpha}_h$  in (6)

**for** every  $i$  such that  $G_i > h - 1$  **do** sample the Pòlya-Gamma data  $\omega_{ih}$  from

$$(\omega_{ih} \mid -) \sim \text{PG}\{1, \boldsymbol{\psi}(\mathbf{x}_i)^\top \boldsymbol{\alpha}_h\}.$$

Given the Pòlya-Gamma augmented data, update  $\boldsymbol{\alpha}_h$  from the full conditional

$$(\boldsymbol{\alpha}_h \mid -) \sim \text{N}_R(\boldsymbol{\mu}_{\boldsymbol{\alpha}_h}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}_h}), \quad \text{having mean and covariance matrix:}$$

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{\alpha}_h} &= \boldsymbol{\Sigma}_{\boldsymbol{\alpha}_h} \{ \boldsymbol{\Psi}_h(\mathbf{x})^\top \boldsymbol{\kappa}_h + \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\alpha}} \} \text{ and } \boldsymbol{\Sigma}_{\boldsymbol{\alpha}_h} = \{ \boldsymbol{\Psi}_h(\mathbf{x})^\top \boldsymbol{\Omega}_h \boldsymbol{\Psi}_h(\mathbf{x}) + \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}^{-1} \}^{-1}, \\ \text{where } \boldsymbol{\Omega}_h &= \text{diag}(\omega_{i1}, \dots, \omega_{i\bar{n}_h}) \text{ and } \boldsymbol{\kappa}_h = (\bar{\zeta}_{i1} - 0.5, \dots, \bar{\zeta}_{i\bar{n}_h} - 0.5)^\top, \text{ with } \bar{\zeta}_{ih} = 1 \\ &\text{if } G_i = h \text{ and } \bar{\zeta}_{ih} = 0 \text{ if } G_i > h. \end{aligned}$$

[3] **for**  $h$  from 1 to  $H$  **do** update each kernel parameter  $\boldsymbol{\beta}_h$  in (2) from

$$(\boldsymbol{\beta}_h \mid -) \sim \text{N}_P(\boldsymbol{\mu}_{\boldsymbol{\beta}_h}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}_h}), \quad \text{having mean and covariance matrix:}$$

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{\beta}_h} &= \boldsymbol{\Sigma}_{\boldsymbol{\beta}_h} \{ \boldsymbol{\Lambda}_h(\mathbf{x})^\top \boldsymbol{\Gamma}_h \mathbf{y}_h + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\beta}} \}, \quad \boldsymbol{\Sigma}_{\boldsymbol{\beta}_h} = \{ \boldsymbol{\Lambda}_h(\mathbf{x})^\top \boldsymbol{\Gamma}_h \boldsymbol{\Lambda}_h(\mathbf{x}) + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} \}^{-1}, \text{ with} \\ \boldsymbol{\Gamma}_h &= \sigma_h^{-2} \mathbf{I}_{n_h \times n_h} \text{ and } \mathbf{y}_h \text{ the } n_h \times 1 \text{ vector containing the response data for all} \\ &\text{subjects with } G_i = h. \end{aligned}$$

[4] **for**  $h$  from 1 to  $H$  **do** sample the precision parameters  $\sigma_h^{-2}$  from

$$(\sigma_h^{-2} \mid -) \sim \text{Ga}[a_\sigma + 0.5 \sum_{i=1}^n \mathbb{1}(G_i = h), b_\sigma + 0.5 \sum_{i: G_i = h} \{y_i - \boldsymbol{\lambda}(\mathbf{x}_i)^\top \boldsymbol{\beta}_h\}^2]$$


---

Generalization to the infinite case can be incorporated leveraging the slice samplers of Walker (2007) and Kalli et al. (2011), which introduce a set of augmented latent variables

allowing each step of the Gibbs sampler to rely on a finite representation. Such strategy effectively slices the infinite mixture model, reducing it to a finite dimensional problem with  $\bar{H}$  mixture components, where  $\bar{H}$  varies stochastically at each step of the chain.

## 3.2 EM algorithm

In several situations — for instance when either  $P$  or  $n$  are large — the Gibbs sampler described in Section 3.1 could face computational bottlenecks. If a point estimate of (2) is the main quantity of interest, one possibility in these high-dimensional problems is to rely on a more efficient procedure specifically designed for this purpose, such as the Expectation Maximization (EM) algorithm (Dempster et al. 1977). The implementation of a simple EM algorithm for a finite representation of model (2) with logit stick-breaking prior (6), greatly benefits from the Pòlya-Gamma data augmentation, which has an explicit form for the expectation and allows analytical maximization within a Gaussian linear regression framework. Note that, although the EM algorithm is commonly used for finding maximum likelihood estimates, it can be easily modified to allow estimation of posterior modes (e.g. Dempster et al. 1977).

The EM method proposed in Algorithm 2 alternates between a maximization step for the parameters  $(\boldsymbol{\alpha}_h, \boldsymbol{\beta}_h, \sigma_h^2)$ ,  $h = 1, \dots, H$ , and an expectation step for the pair of augmented data  $(\boldsymbol{\zeta}_i, \boldsymbol{\omega}_i)$ ,  $i = 1, \dots, n$ , with  $\boldsymbol{\zeta}_i = \{\zeta_{i1} = \mathbb{1}(G_i = 1), \dots, \zeta_{iH} = \mathbb{1}(G_i = H)\}^\top$  the vector of binary indicators denoting membership to a mixture component, and  $\boldsymbol{\omega}_i = (\omega_{i1}, \dots, \omega_{iH})^\top$  the corresponding Pòlya-Gamma augmented data. These steps rely on the complete log-posterior  $l_C(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2 | \mathbf{y}, \boldsymbol{\zeta}, \boldsymbol{\omega})$  which can be defined — up to an additive constant — as

$$\sum_{i=1}^n \log f(y_i, \boldsymbol{\zeta}_i, \boldsymbol{\omega}_i | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2) + \sum_{h=1}^{H-1} \log f(\boldsymbol{\alpha}_h) + \sum_{h=1}^H \log f(\boldsymbol{\beta}_h) + \sum_{h=1}^H \log f(\sigma_h^2), \quad (10)$$

where  $f(y_i, \boldsymbol{\zeta}_i, \boldsymbol{\omega}_i | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$  is the contribution of statistical unit  $i$  to the complete likelihood, whereas  $f(\boldsymbol{\alpha}_h)$ ,  $f(\boldsymbol{\beta}_h)$ , and  $f(\sigma_h^2)$  are the prior density functions for the parameters characterizing our dependent mixture model. Working on  $f(y_i, \boldsymbol{\zeta}_i, \boldsymbol{\omega}_i | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$  has relevant benefits. In fact, exploiting representations (3)–(4), and results in Theorem 1 from Polson et al. (2013), the quantity  $\log f(y_i, \boldsymbol{\zeta}_i, \boldsymbol{\omega}_i | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ , can be factorized as

$$\log f(y_i, \boldsymbol{\zeta}_i, \boldsymbol{\omega}_i | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2) = \log f(y_i | \boldsymbol{\zeta}_i, \boldsymbol{\beta}, \boldsymbol{\sigma}^2) + \log f(\boldsymbol{\zeta}_i | \boldsymbol{\alpha}) + \log f(\boldsymbol{\omega}_i | \boldsymbol{\alpha}),$$

---

**Algorithm 2:** Steps of the EM algorithm for dependent finite mixture of Gaussians

---

Let  $\gamma^{(t)} = (\alpha^{(t)}, \beta^{(t)}, \sigma^{2(t)})$  denote the values of the parameters at iteration  $t$ .

**[1] Expectation Step:** Exploiting results in (11), the expectation of (10) with respect to the augmented data  $(\zeta_i, \omega_i)$ ,  $i = 1, \dots, n$ , can be simply obtained by plugging in  $\hat{\zeta}_i = E(\zeta_i | y_i, \beta^{(t)}, \sigma^{2(t)})$  and  $\hat{\omega}_i = E(\omega_i | \hat{\zeta}_i, \alpha^{(t)})$  in (11). Therefore:  
**for**  $i$  *from* 1 *to*  $n$  **do** compute  $\hat{\zeta}_i$  and  $\hat{\omega}_i$

**for**  $h$  *from* 1 *to*  $H$  **do** compute  $\hat{\zeta}_{ih}$

$$\hat{\zeta}_{ih} = \frac{\left[ \nu_h^{(t)}(\mathbf{x}_i) \prod_{l=1}^{h-1} \{1 - \nu_l^{(t)}(\mathbf{x}_i)\} \right] \frac{1}{\sigma_h^{(t)}} \phi \left\{ \frac{y_i - \lambda(\mathbf{x}_i)^\top \beta_h^{(t)}}{\sigma_h^{(t)}} \right\}}{\sum_{q=1}^H \left[ \nu_q^{(t)}(\mathbf{x}_i) \prod_{l=1}^{q-1} \{1 - \nu_l^{(t)}(\mathbf{x}_i)\} \right] \frac{1}{\sigma_q^{(t)}} \phi \left\{ \frac{y_i - \lambda(\mathbf{x}_i)^\top \beta_q^{(t)}}{\sigma_q^{(t)}} \right\}}.$$

**for**  $h$  *from* 1 *to*  $H$  **do** compute  $\hat{\omega}_{ih}$ , exploiting results in Polson et al. (2013)

$$\hat{\omega}_{ih} = \{2\psi(\mathbf{x}_i)^\top \alpha_h^{(t)}\}^{-1} \tanh \{0.5\psi(\mathbf{x}_i)^\top \alpha_h^{(t)}\} \sum_{l=h}^H \hat{\zeta}_{il}.$$

**[2] Maximization Step:** To maximize the expected complete log-posterior

$E_{(\zeta, \omega | \gamma^{(t)})} \{l_C(\alpha, \beta, \sigma^2 | \mathbf{y}, \zeta, \omega)\} = l_C(\alpha, \beta, \sigma^2 | \mathbf{y}, \hat{\zeta}, \hat{\omega})$ , note that according to (10) and (11), the modes  $\alpha^{(t+1)}$  and  $(\beta^{(t+1)}, \sigma^{2(t+1)})$  can be obtained separately as follow:  
**for**  $h$  *from* 1 *to*  $H - 1$  **do** compute  $\alpha_h^{(t+1)}$ . Since  $\alpha_h$  has Normal prior, and provided that the second term in (11) is based on Gaussian kernels,  $\alpha_h^{(t+1)}$  coincides with the mean of a full conditional Gaussian, similar to the one in step [2] of Algorithm 1.

$$\alpha_h^{(t+1)} = \{\Psi(\mathbf{x})^\top \text{diag}(\hat{\omega}_{1h}, \dots, \hat{\omega}_{nh}) \Psi(\mathbf{x}) + \Sigma_\alpha^{-1}\}^{-1} \{\Psi(\mathbf{x})^\top (\hat{\kappa}_{1h}, \dots, \hat{\kappa}_{nh})^\top + \Sigma_\alpha^{-1} \mu_\alpha\}$$

**for**  $h$  *from* 1 *to*  $H$  **do** compute  $\beta_h^{(t+1)}$  and  $\sigma_h^{2(t+1)}$ . Following a similar reasoning for the computation of  $\alpha_h^{(t+1)}$ , and recalling the Gaussian and Inverse-Gamma priors for  $\beta_h^{(t+1)}$  and  $\sigma_h^{2(t+1)}$ , respectively, this maximization relies on a simple adaptation of steps [3] and [4] in Algorithm 1 to the EM setting, providing:

$$\begin{aligned} \beta_h^{(t+1)} &= \{\Lambda(\mathbf{x})^\top \hat{\Gamma}_h^{(t)} \Lambda(\mathbf{x}) + \Sigma_\beta^{-1}\}^{-1} \{\Lambda(\mathbf{x})^\top \hat{\Gamma}_h^{(t)} \mathbf{y} + \Sigma_\beta^{-1} \mu_\beta\}, \\ \sigma_h^{-2(t+1)} &= \max\{0, [a_\sigma + 0.5 \sum_{i=1}^n \hat{\zeta}_{ih} - 1][b_\sigma + 0.5 \sum_{i=1}^n \hat{\zeta}_{ih} \{y_i - \lambda(\mathbf{x}_i)^\top \beta_h^{(t)}\}^2]^{-1}\} \\ &\text{with } \hat{\Gamma}_h^{(t)} = \sigma_h^{-2(t)} \text{diag}(\hat{\zeta}_{1h}, \dots, \hat{\zeta}_{nh}). \end{aligned}$$


---

where

$$\begin{aligned}
\log f(y_i \mid \boldsymbol{\zeta}_i, \boldsymbol{\beta}, \boldsymbol{\sigma}^2) &\propto \sum_{h=1}^H \zeta_{ih} \left[ -\frac{\{y_i - \boldsymbol{\lambda}(\mathbf{x}_i)^\top \boldsymbol{\beta}_h\}^2}{2\sigma_h^2} - \frac{1}{2} \log(\sigma_h^2) \right], \\
\log f(\boldsymbol{\zeta}_i \mid \boldsymbol{\alpha}) &\propto \sum_{h=1}^{H-1} \left\{ \bar{\kappa}_{ih} \boldsymbol{\psi}(\mathbf{x}_i)^\top \boldsymbol{\alpha}_h + \log E_{\omega_{ih}} \left( \exp \left[ \frac{-\omega_{ih} \{\boldsymbol{\psi}(\mathbf{x}_i)^\top \boldsymbol{\alpha}_h\}^2}{2} \right] \right) \right\}, \\
\log f(\boldsymbol{\omega}_i \mid \boldsymbol{\alpha}) &\propto \sum_{h=1}^{H-1} \left\{ -\frac{\omega_{ih} \{\boldsymbol{\psi}(\mathbf{x}_i)^\top \boldsymbol{\alpha}_h\}^2}{2} - \log E_{\omega_{ih}} \left( \exp \left[ \frac{-\omega_{ih} \{\boldsymbol{\psi}(\mathbf{x}_i)^\top \boldsymbol{\alpha}_h\}^2}{2} \right] \right) \right\},
\end{aligned}$$

with  $\bar{\kappa}_{ih} = \zeta_{ih} - 0.5 \sum_{l=h}^H \zeta_{il}$ , according to the continuation-ratio representation of the logit stick-breaking discussed in equation (4) and Figure 1. Hence, the contribution  $f(y_i, \boldsymbol{\zeta}_i, \boldsymbol{\omega}_i \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$  of statistical unit  $i$  to the complete likelihood is proportional to

$$\sum_{h=1}^H \zeta_{ih} \left[ -\frac{\{y_i - \boldsymbol{\lambda}(\mathbf{x}_i)^\top \boldsymbol{\beta}_h\}^2}{2\sigma_h^2} - \frac{1}{2} \log(\sigma_h^2) \right] + \sum_{h=1}^{H-1} \left\{ \bar{\kappa}_{ih} \boldsymbol{\psi}(\mathbf{x}_i)^\top \boldsymbol{\alpha}_h - \frac{\omega_{ih} \{\boldsymbol{\psi}(\mathbf{x}_i)^\top \boldsymbol{\alpha}_h\}^2}{2} \right\}, \quad (11)$$

where both terms in equation (11) are linear in the augmented data  $(\boldsymbol{\zeta}_i, \boldsymbol{\omega}_i)$ , and represent the sum of Gaussian kernels. The linearity property greatly simplifies computations in the expectation step for the complete log-posterior in equation (10), whereas the Gaussian structure allows simple maximizations. Since the joint maximization of the expected complete log-posterior with respect to  $(\boldsymbol{\beta}, \boldsymbol{\sigma}^2)$  is intractable, we rely on a conditional maximization procedure in the last step of Algorithm 2, which provides closed form solutions. This approach is referred as Expectation Conditional Maximization (Meng & Rubin 1993), and still preserves the monotonicity of the EM sequence. Moreover, by exploiting the Pòlya-Gamma data augmentation, Algorithm 2 avoids approximations via iterated weighted least squares (Nelder & Wedderburn 1972), thereby providing a novel and alternative estimation method based on exact maximization steps.

### 3.3 Variational inference

Section 3.2 provides a scalable procedure for estimation of posterior modes in large-scale problems. However, an appealing aspect of the Bayesian approach is in allowing uncertainty quantification via inference on the entire posterior distribution. As discussed in Section 3.2 the Gibbs sampler represents an appealing procedure which converges to the exact posterior, but faces computational bottlenecks in large-scale applications. This motivates



the development of scalable variational methods for approximate and tractable Bayesian inference (e.g. Bishop 2006, Ch. 10).

In developing a variational Bayes approach for scalable and approximate inference on the dependent mixture model in (2), with logit stick-breaking (6), we first draw a relevant connection with the Bayesian hierarchical mixtures of experts (Bishop & Svensén 2003), and then adapt their variational Bayes algorithm to our construction. In fact, it is easy to show that our density regression model represents a special case of Bayesian hierarchical mixtures of experts, having Gaussian linear regressions at each expert node, and a specific decision tree structure for the gating nodes  $z_{ih} \sim \text{Bern}\{\nu_h(\mathbf{x}_i)\}$ ,  $h = 1, \dots, H - 1$ , for each  $i = 1, \dots, n$ , induced by the logit stick-breaking representation in Figure 1. Note that, differently from the algorithms in Section 3.1 and 3.2, we focus here on the reparameterization  $\mathbb{1}(G_i = h) = \zeta_{ih} = z_{ih} \prod_{l=1}^{h-1} (1 - z_{il})$  to maintain the connection with Bishop & Svensén (2003) and provide simpler analytical derivations.

Leveraging the above results — and using a similar notation as in Bishop & Svensén (2003) — our goal is to find a suitable variational distribution  $q(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{z})$  that best approximates the joint posterior distribution  $f(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{z} \mid \mathbf{y})$ , while maintaining tractable computations. This goal is accomplished by minimizing the Kullback-Leibler divergence  $\text{KL}\{q(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{z}) \parallel f(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{z} \mid \mathbf{y})\}$  between the variational distribution and the full posterior, or alternatively by maximizing the lower bound  $\mathcal{L}\{q(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{z})\}$  of the marginal log-density  $\log f(\mathbf{y} \mid \mathbf{x})$ , provided that

$$\log f(\mathbf{y} \mid \mathbf{x}) = \mathcal{L}\{q(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{z})\} + \text{KL}\{q(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{z}) \parallel f(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{z} \mid \mathbf{y})\},$$

where the lower bound  $\mathcal{L}\{q(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{z})\}$  to be maximized, is defined as

$$\mathcal{L}\{q(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{z})\} = \sum_{\mathbf{z}} \int q(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{z}) \log \left\{ \frac{f(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2) f(\mathbf{z} \mid \boldsymbol{\alpha}) f(\boldsymbol{\alpha}) f(\boldsymbol{\beta}) f(\boldsymbol{\sigma}^2)}{q(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{z})} \right\} d(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\sigma}^2).$$

As discussed in Bishop & Svensén (2003), without further restrictions, the maximization of  $\mathcal{L}\{q(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{z})\}$  with respect to  $q(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{z})$  is analytically intractable due to the logistic transformation in (6), which breaks the conditional conjugacy of the model. This problem can be addressed by an additional variational technique, described in Jaakkola & Jordan (2000) and Bishop & Svensén (2003), which provides a lower bound for

$$\nu_h(\mathbf{x}_i)^{z_{ih}} \{1 - \nu_h(\mathbf{x}_i)\}^{1-z_{ih}} = \exp \{z_{ih} \boldsymbol{\psi}(\mathbf{x}_i)^\top \boldsymbol{\alpha}_h\} [1 + \exp \{\boldsymbol{\psi}(\mathbf{x}_i)^\top \boldsymbol{\alpha}_h\}]^{-1}, \quad (12)$$

by exploiting the inequality

$$\frac{1}{1 + \exp\{\psi(\mathbf{x}_i)^\top \boldsymbol{\alpha}_h\}} \geq \frac{1}{1 + \exp(-\xi_{ih})} \exp\left[-\frac{\psi(\mathbf{x}_i)^\top \boldsymbol{\alpha}_h + \xi_{ih}}{2} - \frac{\tanh(\xi_{ih}/2)}{4\xi_{ih}}\{(\psi(\mathbf{x}_i)^\top \boldsymbol{\alpha}_h)^2 - \xi_{ih}^2\}\right],$$

where the variational parameter  $\xi_{ih}$  is a real number, for  $i = 1, \dots, n$  and  $h = 1, \dots, H-1$ . Therefore, substituting  $[1 + \exp\{\psi(\mathbf{x}_i)^\top \boldsymbol{\alpha}_h\}]^{-1}$  in the right-hand side of (12) with its lower bound, and replacing every  $\nu_h(\mathbf{x}_i)^{z_{ih}}\{1 - \nu_h(\mathbf{x}_i)\}^{1-z_{ih}}$  with these new quantities in  $f(\mathbf{z} \mid \boldsymbol{\alpha})$ , we obtain a further lower bound  $\mathcal{L}_\xi\{q(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{z})\} \leq \mathcal{L}\{q(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{z})\}$ , which preserves conjugacy and is defined as

$$\mathcal{L}_\xi\{q(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{z})\} = \sum_{\mathbf{z}} \int q(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{z}) \log \left\{ \frac{f(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2) f_\xi(\mathbf{z} \mid \boldsymbol{\alpha}) f(\boldsymbol{\alpha}) f(\boldsymbol{\beta}) f(\boldsymbol{\sigma}^2)}{q(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{z})} \right\} d(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\sigma}^2).$$

To conclude the construction of our variational Bayes methodology, we need a functional form for  $q(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{z})$  which leads to tractable computations. In fact, if the variational distribution is left unspecified, the Kullback-Leibler divergence would be minimized by the true posterior, which is intractable. Consistent with this discussion, we consider a mean field approximation — a common procedure in variational Bayes methods — obtaining

$$q(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{z}) = q(\boldsymbol{\alpha})q(\boldsymbol{\beta})q(\boldsymbol{\sigma}^2)q(\mathbf{z}) = \prod_{h=1}^{H-1} q(\boldsymbol{\alpha}_h) \prod_{h=1}^H q(\boldsymbol{\beta}_h) \prod_{h=1}^{H-1} q(\boldsymbol{\sigma}_h^2) \prod_{h=1}^{H-1} \prod_{i=1}^n q(z_{ih}).$$

This factorization greatly simplifies the functional form of  $\mathcal{L}_\xi\{q(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{z})\}$ , and allows tractable maximization of the variational distributions for each block of parameters in turn. In particular, following for example Bishop (2006, Ch. 10), the optimal solutions for the variational distribution — under the above mean field approximation — are provided by the following equations

$$\begin{aligned} \log q^*(z_{ih}) &= E_{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{z}_{(-h)}}[\log\{f(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2) f_\xi(\mathbf{z} \mid \boldsymbol{\alpha})\}] + c_{z_{ih}}, \quad i = 1, \dots, n, \\ \log q^*(\boldsymbol{\alpha}_h) &= E_{\mathbf{z}}[\log\{f_\xi(\mathbf{z} \mid \boldsymbol{\alpha}) f(\boldsymbol{\alpha}_h)\}] + c_{\boldsymbol{\alpha}_h}, \\ \log q^*(\boldsymbol{\beta}_h) &= E_{\mathbf{z}, \boldsymbol{\sigma}^2}[\log\{f(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2) f(\boldsymbol{\beta}_h)\}] + c_{\boldsymbol{\beta}_h}, \\ \log q^*(\boldsymbol{\sigma}_h^2) &= E_{\mathbf{z}, \boldsymbol{\beta}}[\log\{f(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2) f(\boldsymbol{\sigma}_h^2)\}] + c_{\boldsymbol{\sigma}_h^2}, \end{aligned}$$

for every  $h = 1, \dots, H$ , where  $\mathbf{z}_{(-h)}$  denotes the vector of gating nodes without considering the  $h$  one, whereas  $c_{z_{ih}}$ ,  $c_{\boldsymbol{\alpha}_h}$ ,  $c_{\boldsymbol{\beta}_h}$  and  $c_{\boldsymbol{\sigma}_h^2}$  are additive constants with respect to the argument in the corresponding variational distribution. Estimation of the variational parameters  $\xi_{ih}$ ,

---

**Algorithm 3:** Steps of the VB algorithm for dependent finite mixture of Gaussians

---

Let  $q^{(t)}(\cdot)$  denote the generic variational distribution at iteration  $t$ .

[1] **for**  $i$  *from* 1 *to*  $n$  **do**

**for**  $h$  *from* 1 *to*  $H - 1$  **do**

        It can be easily shown that the optimal solution for the variational distribution  $q(z_{ih})$  of each  $z_{ih}$  coincides with  $q^{*(t)}(z_{ih}) = \text{Bern}(\rho_{ih})$ , where

$$\text{logit}(\rho_{ih}) = \boldsymbol{\psi}(\mathbf{x}_i)^\top \mathbf{E}(\boldsymbol{\alpha}_h) + \sum_{l=h}^H \zeta_{il}^{(h)} \left[ \frac{\mathbf{E}(\log \sigma_l^{-2})}{2} - \frac{\mathbf{E}(\sigma_l^{-2})}{2} \mathbf{E}\{(y_i - \boldsymbol{\lambda}(\mathbf{x})^\top \boldsymbol{\beta}_l)^2\} \right],$$

        where the expectations are taken with the respect to the current variational distributions for the other parameters, whereas  $\zeta_{il}^{(h)} = \prod_{r=1}^{l-1} (1 - \rho_{ir})$  if  $l = h$ , and  $\zeta_{il}^{(h)} = -\rho_{il} \prod_{r=1, r \neq h}^{l-1} (1 - \rho_{ir})$  otherwise. Note also that  $\rho_{iH} = 1$ .

[2] **for**  $h$  *from* 1 *to*  $H - 1$  **do**

    Following Jaakkola & Jordan (2000), the optimal solution  $q^{*(t)}(\boldsymbol{\alpha}_h)$  for the variational distribution of each  $\boldsymbol{\alpha}_h$  is easily available as a Gaussian distribution

$$q^{*(t)}(\boldsymbol{\alpha}_h) = \text{N}_R[\{\boldsymbol{\Psi}(\mathbf{x})^\top \mathbf{V}_h \boldsymbol{\Psi}(\mathbf{x}) + \boldsymbol{\Sigma}_\alpha^{-1}\}^{-1} \{\boldsymbol{\Psi}(\mathbf{x})^\top \bar{\boldsymbol{\rho}}_h + \boldsymbol{\Sigma}_\alpha^{-1} \boldsymbol{\mu}_\alpha\}, \{\boldsymbol{\Psi}(\mathbf{x})^\top \mathbf{V}_h \boldsymbol{\Psi}(\mathbf{x}) + \boldsymbol{\Sigma}_\alpha^{-1}\}^{-1}]$$

    with  $\mathbf{V}_h = \text{diag}\left(\frac{\tanh(\xi_{1h}^{*(t)}/2)}{2\xi_{1h}^{*(t)}}, \dots, \frac{\tanh(\xi_{nh}^{*(t)}/2)}{2\xi_{nh}^{*(t)}}\right)$  and

$$\bar{\boldsymbol{\rho}}_h = (\rho_{1h} - 1/2, \dots, \rho_{nh} - 1/2).$$

[3] **for**  $h$  *from* 1 *to*  $H$  **do**

    Update the optimal solution for the variational distributions of  $\boldsymbol{\beta}_h$  and  $\sigma_h^2$  via

$$q^{*(t)}(\boldsymbol{\beta}_h) = \text{N}_P[\{\boldsymbol{\Lambda}(\mathbf{x})^\top \boldsymbol{\Gamma}_h \boldsymbol{\Lambda}(\mathbf{x}) + \boldsymbol{\Sigma}_\beta^{-1}\}^{-1} \{\boldsymbol{\Lambda}(\mathbf{x})^\top \boldsymbol{\Gamma}_h \mathbf{y} + \boldsymbol{\Sigma}_\beta^{-1} \boldsymbol{\mu}_\beta\}, \{\boldsymbol{\Lambda}(\mathbf{x})^\top \boldsymbol{\Gamma}_h \boldsymbol{\Lambda}(\mathbf{x}) + \boldsymbol{\Sigma}_\beta^{-1}\}^{-1}]$$

$$q^{*(t)}(\sigma_h^{-2}) = \text{Ga}[a_\sigma + 0.5 \sum_{i=1}^n \mathbf{E}(\zeta_{ih}), b_\sigma + 0.5 \sum_{i=1}^n \mathbf{E}(\zeta_{ih}) \mathbf{E}\{y_i - \boldsymbol{\lambda}(\mathbf{x}_i)^\top \boldsymbol{\beta}_h\}^2]$$

    with  $\boldsymbol{\Gamma}_h = \mathbf{E}(\sigma_h^{-2}) \text{diag}\{\mathbf{E}(\zeta_{1h}), \dots, \mathbf{E}(\zeta_{nh})\}$ .

[4] **for**  $i$  *from* 1 *to*  $n$  **do**

**for**  $h$  *from* 1 *to*  $H - 1$  **do**

        Following Bishop & Svensén (2003) the maximum of the variational parameter can be easily obtained via  $\xi_{ih}^{*(t)} = \boldsymbol{\psi}(\mathbf{x}_i)^\top \mathbf{E}(\boldsymbol{\alpha}_h \boldsymbol{\alpha}_h^\top) \boldsymbol{\psi}(\mathbf{x}_i)$ .

for every  $i = 1, \dots, n$  and  $h = 1, \dots, H - 1$  at each step is instead straightforward, since it requires the maximization of the expected lower bound provided by Jaakkola & Jordan (2000), where the expectation is taken with respect to the variational distribution of  $\boldsymbol{\alpha}_h$ . Since each expectation in the above equations is evaluated with respect to the variational distribution of the other parameters, we consider an iterative procedure — described in Algorithm 3 — which maximizes the variational distribution of each parameter based on the current estimate for the remaining ones (e.g. Bishop 2006, Ch. 10). This procedure generates a monotonic sequence of  $\mathcal{L}_{\xi}\{q(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{z})\}$ , which ensures convergence to a local joint maximum. Moreover, as shown in Algorithm 3, the normalizing constants in the above equations do not need be computed explicitly at every step, since the variational approximation we consider produces kernels of well known distributions under our Bayesian density regression settings.

## 4 The Old Faithful Geyser dataset

We evaluate the performance of the three computational methods developed in Section 3, focusing on the Old Faithful Geyser dataset (e.g. Azzalini & Bowman 1990), which is available in the R library `MASS`. Data consists of  $n = 299$  measurements  $(x_i, y_i)$ ,  $i = 1, \dots, n$ , on the behavior of the Old Faithful Geyser, collected during August 1985 in the Yellowstone National Park. The covariate  $x_i$  represents the waiting time between consecutive eruptions  $i - 1$  and  $i$ , whereas the response  $y_i$  measures the duration of eruption  $i$ . Although statistical analyses of these data focused on modeling the temporal dependence in the two time series  $\{x_i : i = 1, \dots, n\}$  and  $\{y_i : i = 1, \dots, n\}$ , our main goal is to assess performance of the proposed statistical model and the associated computational methods. Therefore, we focus on studying how the distribution of the eruption’s durations  $y_i$  changes as a function of the corresponding waiting times  $x_i$ .

As discussed in Azzalini & Bowman (1990), and consistent with Figure 3, for small waiting times — below  $\approx 68$  minutes — the durations are generated by a single-mode distribution centered around  $\approx 4.5$  minutes. When instead the waiting time increases — exceeding the  $\approx 68$  minutes — the durations are characterized by a bimodal distribution which assigns growing mass to the component associated with low durations of  $\approx 2$  min-

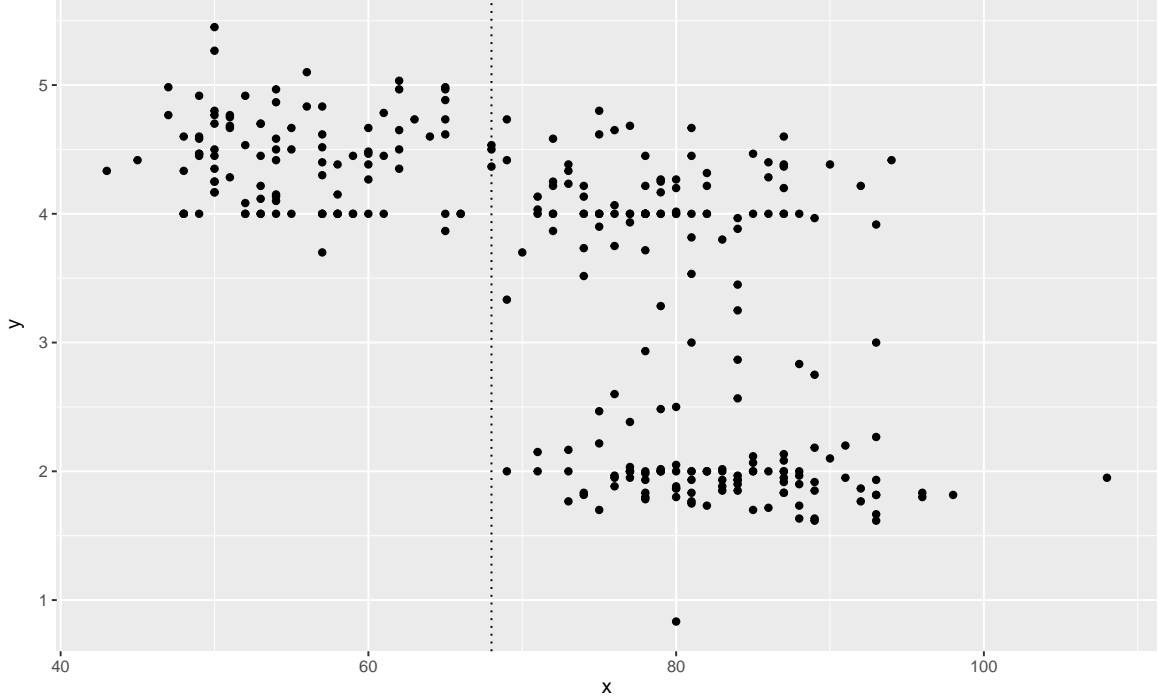


Figure 3: Scatterplot of the waiting time  $x_i$  against the duration time  $y_i$ , expressed in minutes. A vertical dotted line is drawn at  $x = 68$ .

utes. This type of association motivates our dependent mixture of Gaussians described in equation (2), with logit stick-breaking prior (6) for the mixing probabilities. In particular, we reasonably expect two Gaussian mixture components with expectations  $\approx 4.5$  and  $\approx 2$ , respectively, not changing with the predictor. The probability assigned to such components is instead varying with the waiting times. In fact, for values of the predictor less than  $\approx 68$ , almost all the mass is assigned to the first component, whereas the second is increasingly populated as the waiting time increases — after exceeding  $\approx 68$  minutes.

Consistent with the above discussion we consider predictor-independent Gaussian mixture components in (2) by letting  $P = 1$ , with  $\lambda_1(x_i) = 1$ , for every  $i = 1, \dots, n$ , and rely on a piecewise continuous linear function to characterize changes in the logit stick-breaking prior with the waiting times  $x_i$ , obtaining

$$\text{logit}\{\nu_h(x_i)\} = \eta_h(x_i) = \boldsymbol{\psi}(x_i)^\top \boldsymbol{\alpha}_h = \alpha_{1h} + \alpha_{2h}\psi_2(x_i) + \alpha_{3h}\psi_3(x_i), \quad (13)$$

for every  $h = 1, \dots, H - 1$ , with  $\psi_2(x_i) = x_i\mathbb{1}(x_i \leq 68) + 68\mathbb{1}(x_i > 68)$  and  $\psi_3(x_i) = (x_i - 68)\mathbb{1}(x_i > 68)$ . Equation (13) provides a mathematical representation of a piecewise

continuous linear function with a knot in 68. Therefore  $\alpha_{1h}$  denotes the value of the linear predictor when  $x_i = 0$ , whereas  $\alpha_{2h}$  and  $\alpha_{3h}$  are the slopes of the two linear functions before and after the knot, respectively, for every stick-breaking weight  $h = 1, \dots, H - 1$ .

We perform Bayesian posterior inference for the above model, under the three computational methods developed in Section 3, setting the hyperparameters at  $\mu_\beta = 0$ ,  $\sigma_\beta^2 = 10$ ,  $\mu_\alpha = (0, 0, 0)^\top$ ,  $\Sigma_\alpha = 10I_{3 \times 3}$  and  $a_\sigma = b_\sigma = 2$ . For the total number of mixture components we consider a conservative choice  $H = 5$ , and allow the shrinkage induced by the stick-breaking prior to adaptively delete redundant components not required to characterize the data. In providing posterior inference under the Gibbs sampling algorithm described in Section 3.1, we rely on 30,000 iterations, after discarding the first 5,000 as a burn-in. The analysis of the traceplots for the functionals of interest — displayed in Figure 4 — showed that this choice is sufficient for good convergence. The EM algorithm and the variational Bayes procedures discussed in Sections 3.2 and 3.3, respectively, are instead run until convergence to a modal solution. Since such modes could be only local, we run both algorithms for different initial values, and consider the solutions having the highest values for the log-posterior and for the lower bound of the marginal density, respectively. We also controlled the monotonicity of the sequences for these quantities, in order to further validate the correctness of our derivations. These algorithms were implemented through R and C++ coding, and made available online at <https://github.com/tommasorigon/DLSBP>. In this application, the EM and the variational Bayes algorithms reach convergence in about 0.625 and 2.887 seconds, respectively, whereas the Gibbs sampler requires 58 seconds, using a standard notebook with a Intel Core i7 processor.

Consistent with some goals motivating the analysis of the Old Faithful Geyser dataset in Azzalini & Bowman (1990), Figure 4 provides inference on two key functionals of  $f(y | x)$ , covering the conditional expectation  $E(y | x) = \sum_{h=1}^H \pi_h(x) \beta_{1h}$  and the conditional probability of a long eruption  $\text{pr}(y > 3 | x) = \sum_{h=1}^H \pi_h(x) [1 - \Phi\{(3 - \beta_{1h})\sigma_h^{-1}\}]$  — obtained under the developed algorithms. Both quantities are available as functionals of the parameters, and therefore can be easily computed from the output of the three computational methods. The Gibbs sampler and the variational Bayes provide the entire posterior distribution for these quantities, thereby allowing uncertainty quantification. The EM algorithm allows

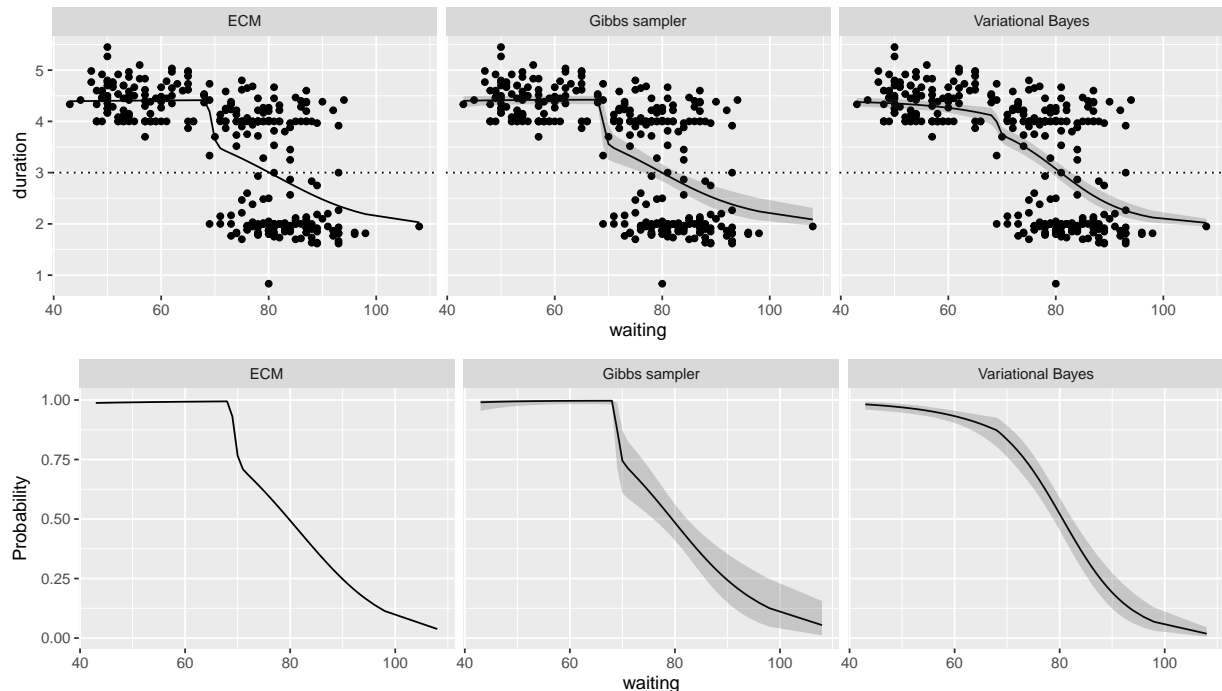


Figure 4: Posterior mean — estimated under the three different proposed algorithms — for two functionals of  $f(y | x)$  of interest, covering  $E(y | x)$  (upper panels) and  $\text{pr}(y > 3 | x)$  (bottom panels). The black dots in the upper panels denote the observed data, whereas the gray areas denote the 0.95 credibility intervals. These quantities are not available for the EM algorithm, since it provides only a point estimate of  $f(y | x)$ .

instead only point estimation via plug-in of the posterior modes for the model parameters. Results in Figure 4 confirm the initial graphical analyses of the observed data, while showing relevant similarities across the different computational methods, which empirically guarantee the goodness of our algorithms. As expected, the point estimates provided by the EM algorithm match the posterior mean from the Gibbs sampler. The results from the variational Bayes are instead characterized by an higher level of smoothing, which however does not substantially affect the final conclusions. This is a reasonable result, given that the variational Bayes provides only a mean field approximation of the posterior distribution.

We obtain similar results in the goodness-of-fit assessments displayed in Figure 5. Under the Gibbs sampler and the variational Bayes, these checks can formally proceed via the conditional posterior predictive density (e.g. Gelman et al. 2013, Chap. 6), obtained by

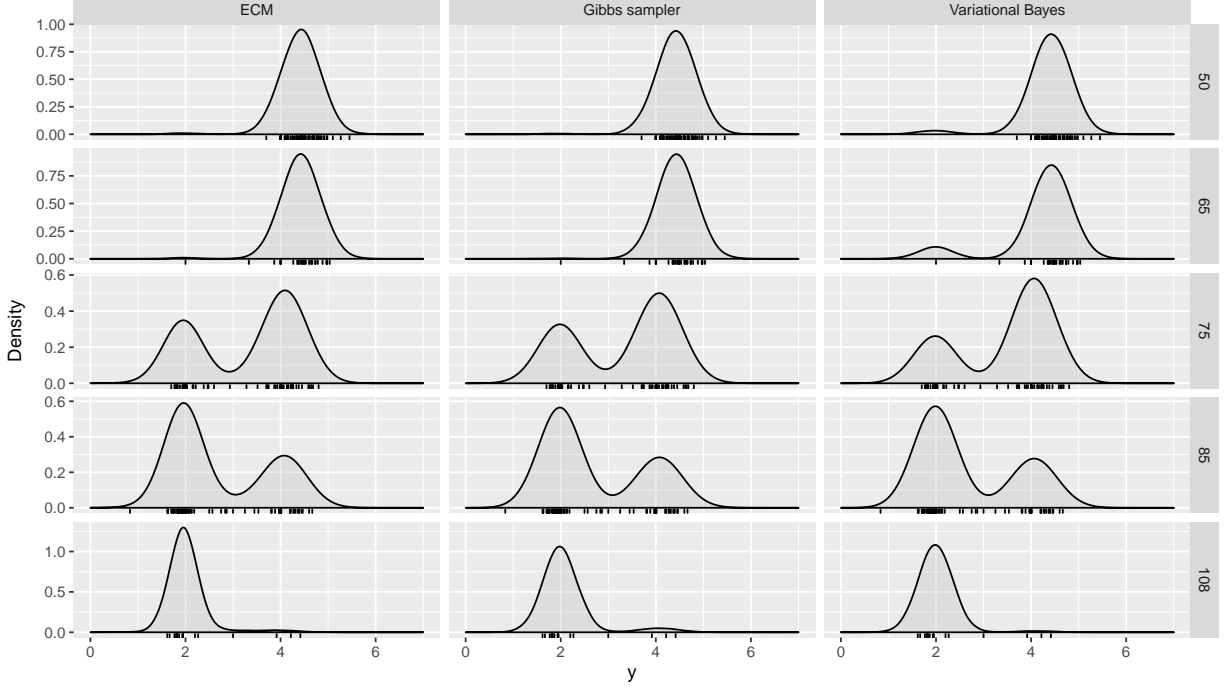


Figure 5: For selected values of  $x \in (50, 65, 75, 85, 108)$ , graphical representation of the posterior predictive density for  $y$  given  $x$ , obtained from the Gibbs sampler and the variational Bayes. Since the EM provides only a mode for the conditional density, we consider a graphical representation of the plug-in estimate for  $f(y | x)$ . The lines in the horizontal axes represent the observations  $y_i$ , having  $x_i$  in the intervals  $(-\infty, 60)$ ,  $[60, 70)$ ,  $[70, 80)$ ,  $[80, 90)$ ,  $[90, +\infty)$ , respectively.

marginalizing out the model parameters in (2) and (6) with the respect to their posterior distribution. Leveraging the hierarchical representation (3) of our model, the conditional posterior predictive density can be easily obtained via Monte Carlo integration. The EM algorithm returns instead only a point estimate of the model parameters, and therefore a proper posterior predictive distribution cannot be computed. However, a goodness-of-fit assessment can still be performed via a plug-in of the parameters estimates in equations (2) and (6). Although this procedure does not incorporate posterior uncertainty on the model parameters, the conditional densities obtained via a plug-in of the EM estimates produce results which are indistinguishable from the those obtained from the Gibbs sampler and the variational Bayes. Similar conclusions are obtained when comparing the posterior



predictive densities induced by the Gibbs sampler and the variational Bayes method, further confirming the goodness of our algorithms, and the flexibility of the dependent mixture of Gaussians in modeling the process underlying the observed data.

## 5 Discussion

The focus of this paper has been on providing a flexible class of Bayesian density regression models based on a dependent mixture of Gaussian distributions, with mixing probabilities changing with the predictor via a logit stick-breaking prior. This class of models is constructed to have theoretical support, and to facilitate implementation of tractable procedures for posterior inference. In fact, we discuss a constructive representation of the assignment process to the mixture components via a continuation-ratio logistic regression, which allows derivation of routine use computational methods. The Gibbs sampler and the Expectation Maximization algorithms rely on the Pòlya-Gamma data augmentation for Bayesian logistic regression, whereas the variational Bayes greatly benefits from the connection between our statistical model and the Bayesian hierarchical mixtures of experts. These methods are compared and empirically evaluated in an application to the Old Faithful Geyser dataset, providing good results.

Although our dependent mixture of Gaussians provides a flexible representation, it is worth considering extensions to more general kernels. For example, all our algorithms can be easily adapted to predictor-independent kernels coming from an exponential family, when conjugate priors for their parameters are used. Similar derivations are also possible for more general predictor-dependent kernels having a generalized linear model representation — provided that conjugate priors for the coefficients can be found (e.g. Chen & Ibrahim 2003). Theory and computational steps associated with the logit stick-breaking prior on the mixing probabilities are instead general and remain valid regardless the kernel choice.

## References

Aitchison, J. & Shen, S. M. (1980), ‘Logistic-Normal Distributions: Some Properties and Uses’, *Biometrika* **67**, 262–272.

- Amemiya, T. (1981), ‘Qualitative Response Models: A Survey’, *Journal of Economic Literature* **19**, 1483–1536.
- Antoniano-Villalobos, I., Wade, S. & Walker, S. (2014), ‘A Bayesian Nonparametric Regression Model with Normalized Weights: A Study of Hippocampal Atrophy in Alzheimer’s Disease’, *Journal of the American Statistical Association* **109**, 477–490.
- Azzalini, A. & Bowman, A. W. (1990), ‘A Look at Some Data on the Old Faithful Geyser’, *Journal of the Royal Statistical Society, Series C* **39**, 357–365.
- Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, Springer.
- Bishop, C. M. & Svensén, M. (2003), ‘Bayesian Hierarchical Mixture of Experts’, in *Proceedings of the nineteenth conference of uncertainty of artificial intelligence*, pp. 1-6.
- Caron, F., Davy, M., Doucet, A., Duflos, E. & Vanheeghe, P. (2006), ‘Bayesian Inference for Dynamic Models with Dirichlet Process Mixtures’, in *9th IEEE International Conference on Information Fusion*, pp. 1–8.
- Chen, M.-H. & Ibrahim, J. G. (2003), ‘Conjugate Priors for Generalized Linear Models’, *Statistica Sinica* **13**, 461–476.
- Choi, H. M. & Hobert, J. P. (2013), ‘The Polya-Gamma Gibbs Sampler for Bayesian Logistic Regression is Uniformly Ergodic’, *Electronic Journal of Statistics* **7**, 2054–2064.
- De Iorio, M., Müller, P., Rosner, G. L. & MacEachern, S. N. (2004), ‘An ANOVA Model for Dependent Random Measures’, *Journal of the American Statistical Association* **99**, 205–215.
- De la Cruz-Mesía, R., Quintana, F. A. & Müller, P. (2007), ‘Semiparametric Bayesian Classification with Longitudinal Markers’, *Journal of the Royal Statistical Society: Series C* **56**, 119–137.
- Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), ‘Maximum Likelihood from Incomplete Data via the EM Algorithm’, *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

- Dunson, D. B. & Park, J. H. (2008), ‘Kernel Stick-Breaking Processes’, *Biometrika* **95**, 307–323.
- Escobar, M. D. & West, M. (1995), ‘Bayesian Density Estimation and Inference Using Mixtures’, *Journal of the American Statistical Association* **90**, 577–588.
- Ferguson, T. S. (1973), ‘A Bayesian Analysis of Some Nonparametric Problems’, *The Annals of Statistics* **1**, 209–230.
- Ferguson, T. S. (1974), ‘Prior Distributions on Spaces of Probability Measures’, *The Annals of Statistics* **2**, 615–629.
- Gelfand, A. E., Kottas, A. & MacEachern, S. N. (2005), ‘Bayesian Nonparametric Spatial Modeling with Dirichlet Process Mixing’, *Journal of the American Statistical Association* **100**, 1021–1035.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2013), *Bayesian Data Analysis, third edition*, CRC Press.
- Ghosal, S., Ghosh, J. K. & Ramamoorthi, R. (1999), ‘Posterior Consistency of Dirichlet Mixtures in Density Estimation’, *The Annals of Statistics* **27**, 143–158.
- Ghosal, S. & Van Der Vaart, A. (2007), ‘Posterior Convergence Rates of Dirichlet Mixtures at Smooth Densities’, *The Annals of Statistics* **35**, 697–723.
- Griffin, J. E. & Steel, M. (2006), ‘Order-Based Dependent Dirichlet Processes’, *Journal of the American Statistical Association* **10**, 179–194.
- Griffin, J. E. & Steel, M. F. (2011), ‘Stick-Breaking Autoregressive Processes’, *Journal of Econometrics* **162**, 383–396.
- Gutiérrez, L., Mena, R. H. & Ruggiero, M. (2016), ‘A Time Dependent Bayesian Nonparametric Model for Air Quality Analysis’, *Computational Statistics & Data Analysis* **95**, 161–175.
- Hannah, L. A., Blei, D. M. & Powell, W. B. (2011), ‘Dirichlet Process Mixtures of Generalized Linear Models’, *Journal of Machine Learning Research* **12**, 1923–1953.

- Ishwaran, H. & James, L. F. (2001), ‘Gibbs Sampling Methods for Stick-Breaking Priors’, *Journal of the American Statistical Association* **96**, 161–173.
- Ishwaran, H. & James, L. F. (2002), ‘Approximate Dirichlet Process Computing Finite Normal Mixtures: Smoothing and Prior Information’, *Journal of Computational and Graphical Statistics* **11**, 508–532.
- Jaakkola, T. S. & Jordan, M. I. (2000), ‘Bayesian Parameter Estimation via Variational Methods’, *Statistics and Computing* **10**, 25–37.
- Jordan, M. I. & Jacobs, R. A. (1994), ‘Hierarchical Mixture of Experts and the EM Algorithm’, *Neural Computation* **6**, 181–214.
- Kalli, M., Griffin, J. E. & Walker, S. G. (2011), ‘Slice Sampling Mixture Models’, *Statistics and Computing* **21**, 93–105.
- MacEachern, S. N. (1999), Dependent Nonparametric Processes, in ‘Proceedings of the Bayesian Section’, Alexandria, VA: American Statistical Association, pp. 50–55.
- MacEachern, S. N. (2000), Dependent Dirichlet Processes, Technical report, Department of Statistics, Ohio State University.
- McCullagh, P. & Nelder, J. A. (1989), *Generalized Linear Models*, Chapman and Hall.
- Meng, X.-L. & Rubin, D. B. (1993), ‘Maximum Likelihood Estimation via the ECM Algorithm: a General Framework’, *Biometrika* **80**, 267–278.
- Müller, P., Erkanly, A. & West, M. (1996), ‘Bayesian Curve Fitting Using Multivariate Normal Mixtures’, *Biometrika* **83**, 67–79.
- Müller, P. & Quintana, F. (2010), ‘Random Partition Models with Regression on Covariates’, *Journal of Statistical Planning and Inference* **140**(10), 2801–2808.
- Neal, R. M. (2000), ‘Markov Chain Sampling Methods for Dirichlet Process Mixture Models’, *Journal of Computational and Graphical Statistics* **9**, 249–265.
- Nelder, J. A. & Wedderburn, R. W. M. (1972), ‘Generalized Linear Models’, *Journal of the Royal Statistical Society, Series A* **135**, 370–384.

- Pati, D., Dunson, D. B. & Tokdar, S. T. (2013), ‘Posterior Consistency in Conditional Distribution Estimation’, *Journal of Multivariate Analysis* **116**, 456–472.
- Pitman, J. & Yor, M. (1997), ‘The Two-Parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator’, *The Annals of Probability* **25**, 855–900.
- Polson, N. G., Scott, J. G. & Windle, J. (2013), ‘Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables’, *Journal of the American Statistical Association* **108**, 1339–1349.
- Ren, L., Du, L., Carin, L. & Dunson, D. B. (2011), ‘Logistic Stick-Breaking Process’, *Journal of Machine Learning Research* **12**, 203–239.
- Rodriguez, A. & Dunson, D. B. (2011), ‘Nonparametric Bayesian Models Through Probit Stick-Breaking Processes’, *Bayesian Analysis* **6**, 145–178.
- Sethuraman, J. (1994), ‘A Constructive Definition of Dirichlet Priors’, *Statistica Sinica* **4**, 639–650.
- Tokdar, S. T. (2006), ‘Posterior Consistency of Dirichlet Location-Scale Mixture of Normals in Density Estimation and Regression’, *Sankhyā: The Indian Journal of Statistics* pp. 90–110.
- Tutz, G. (1991), ‘Sequential Models in Categorical Regression’, *Computational Statistics & Data Analysis* **11**, 275–295.
- Wade, S., Dunson, D. B., Petrone, S. & Trippa, L. (2014), ‘Improving Prediction from Dirichlet Process Mixtures via Enrichment’, *Journal of Machine Learning Research* **15**, 1041–1071.
- Walker, S. G. (2007), ‘Sampling the Dirichlet Mixture Model with Slices’, *Communications in Statistics – Simulation and Computation* **36**, 45–54.